話し言葉における引用節・挿入節の自動認定および係り受け解析への応用

浜辺　良二[†]・内元　清貴[††]・河原　達也[†]・井佐原　均[††]

話し言葉の係り受け解析を行なう際の最大の問題は，文境界や引用節・挿入節などの境界が明示されていないことである．本論文では，話し言葉に対して，引用節・挿入節を自動認定するための手法，および自動認定した引用節・挿入節の情報を用いて係り受け解析を改善するための手法を提案する．形態素やポーズの情報などをもとに，SVM を用いたテキストチャンキングによって，引用節・挿入節の始端と終端を決定する．始端を決定する際には，自動推定した係り受けの情報をあわせて利用する．日本語話し言葉コーパス (CSJ) を用いた評価実験により，自動認定した引用節・挿入節の情報を利用することで係り受け解析精度が 77.7% から 78.7% に改善されることを確認し，本手法の有効性を示した．

キーワード：話し言葉，係り受け解析，節境界，引用節，挿入節，機械学習

Detection of Quotations and Inserted Clauses and its Application to Dependency Structure Analysis in Spontaneous Japanese

Ryoji Hamabe[†], Kiyotaka Uchimoto[††], Tatsuya Kawahara[†] and Hitoshi Isahara[††]

Japanese dependency structure is usually represented by relationships between phrasal units called *bunsetsu*s. One of the biggest problems with dependency structure analysis in spontaneous speech is that clause boundaries are ambiguous. This paper describes a method for detecting the boundaries of quotations and inserted clauses and that for improving the dependency accuracy by applying the detected boundaries to dependency structure analysis. The quotations and inserted clauses are determined by using an SVM-based text chunking method that considers information on morphemes, pauses, etc. The information on automatically analyzed dependency structure is also used to detect the beginning of the clauses. Our evaluation experiment using *Corpus of Spontaneous Japanese (CSJ)* showed that the automatically estimated boundaries of quotations and inserted clauses helped to improve the accuracy of dependency structure analysis from 77.7% to 78.7% .

Key Words: *spontaneous Japanese, dependency structure analysis, clause boundary, quotation, inserted clause, machine learning*

[†] 京都大学 情報学研究科, School of Informatics, Kyoto University
[††] 独立行政法人 情報通信研究機構, National Institute of Information and Communications Technology

1 はじめに

係り受け解析は日本語解析の重要な基本技術の一つとして認識されており，これまでに様々な手法が提案されてきた (黒橋, 長尾 1994; 白井, 池原, 横尾, 木村 1995; 藤尾, 松本 1997; 春野, 白井, 大山 1998; 内元, 関根, 井佐原 1999; 内元, 村田, 関根, 井佐原 2000; Kudo and Matsumoto 2000; 工藤, 松本 2002; Matsubara, Murase, Kawaguchi, and Inagaki 2002; 工藤, 松本 2004; Kawahara and Kurohashi 2006; Ohno, Matsubara, Kashioka, Maruyama, and Inagaki 2006)．しかし，そのほとんどは書き言葉を対象としたものであった．これに対し，本研究では，話し言葉，特に『日本語話し言葉コーパス (CSJ) (古井, 前川, 井佐原 2000)』のような長い独話を対象とする．ここで CSJ とは，主に学会講演や模擬講演などの独話を対象に，約 660 時間（約 750 万語）の自発音声を収録した世界最大規模の話し言葉コーパスのことである．このコーパスには音声データだけでなく書き起こしも含まれており，コアと呼ばれる一部の書き起こしには，人手により形態素・係り受け・節境界・引用節・挿入節・談話構造など様々な情報が付与されている．

一般に，話し言葉には特有の現象が見られるため，書き言葉と比べて話し言葉の係り受け解析は難しい．例えば，CSJ を用いた実験によると，話し言葉特有の現象の影響をなくした場合とそうでない場合で，係り受け解析精度に大きな差があることが報告されている (Uchimoto, Hamabe, Maruyama, Takanashi, Kawahara, and Isahara 2006)．特に，引用節・挿入節などの境界が認識されていない場合に係り受け解析精度の低下が著しい．そこで本論文では，引用節・挿入節を自動認定する方法，および，自動認定した引用節・挿入節の情報を係り受け解析に利用する方法を提案し，提案手法により係り受け解析精度が有意に向上することを定量的に示す．

2 話し言葉に特有の現象と係り受け構造

話し言葉には，書き言葉にはない特有の現象が見られる．そして，その話し言葉特有の現象が係り受け解析精度の低下を招くことが多い．本研究では，その中でも節境界が曖昧であるという現象に着目する．そして，本論文では，係り受け解析精度に及ぼす影響の大きさを考慮し，節の中でも，特に，引用節・挿入節と係り受け構造との関係を取り上げる．ここで，節および係り受け構造の定義は CSJ に従うものとする．

以下，2.1 節では，話し言葉における節境界と係り受け構造の定義，および，引用節と挿入節との関係について述べる．次に，2.2 節では，話し言葉特有の現象，特に，節境界が曖昧であるという現象が係り受け解析に及ぼす影響について言及する．そして，2.3 節では，その他の話し言葉特有の現象に関して，本研究での係り受け解析時の扱いについて述べる．

2.1　節境界と係り受け構造の定義および引用節・挿入節との関係

一般に，書き言葉においては，係り受け構造などを付与する単位として，いわゆる「文」を用いることが多い．しかし，自発的な話し言葉を対象とする場合，文は必ずしも自明な単位ではない．そこで CSJ では，より適切な分割単位として，「節」に基づく文の単位が定義されている．節境界としては，次の 3 種類が定義されている．

絶対境界：いわゆる文末表現で，述語の終止形・終助詞・「と文末」など．

強境界　：並列節「ケレドモ」「ガ」「シ」・「ましテ」節・「でしテ」節など．

弱境界　：理由節「カラ」「ノデ」・連用節・引用節・条件節「タラ」「ト」「ナラ」「レバ」など．

そして，これらの節境界を表層表現などに基づいて自動検出した後 (丸山, 柏岡, 熊野, 田中 2003)，文節係り受けを考慮して人手により文境界を特定する (高梨, 丸山, 内元, 井佐原 2003)．上の 3 種類の境界のうち，絶対境界と強境界は基本的に文境界となり，弱境界は機能的に区切れていると判断される箇所のみが文境界となる．引用節と挿入節についてもこのときに認定され，基本的に，引用節の終端は弱境界，挿入節の終端は強境界となる．

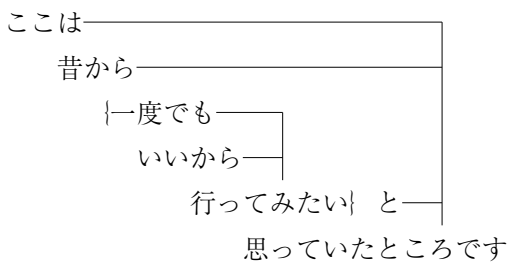
CSJ における係り受け構造は，原則として「京大コーパス」(黒橋 長尾 1997) の付与基準に準拠して付与されており，話し言葉特有の現象に対しては新たな基準が設けられている (内元, 丸山, 高梨, 井佐原 2003)．係り受けは文内で閉じており，引用節・挿入節の内部でも同様に係り受けが閉じている．したがって，文境界が特定されれば，引用節と挿入節の始端は係り受け構造に基づいて特定できる．すなわち，直前の文節が終端より後方に係る文節のうち，最も終端に近いものが引用節の始端となる．そのような文節がない場合は，文頭の文節が始端となる．

話し言葉，特に CSJ のような長い独話における引用節・挿入節の特徴は次の通りである．以下では，引用節・挿入節と係り受け構造との関係の例も示す．

（引用節）

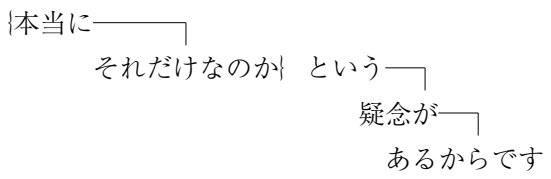
引用節は，主に人の言ったことや思ったことを発話に取り込む際に用いられる．書き言葉では引用節の前後に引用符や読点が付与されるのに対し，発話においては，その境界が明示されることはない．以下の例文 1 では，〦 内が引用節に相当し，「昔から」が引用節の後方に係るため「一度でも」が始端となる．

（例文 1）



CSJ では，引用節の終端の文節に「引用節」というラベルが付与されている．本研究では，それに加え「トイウ節」のラベルの付いたものも引用節として扱う．トイウ節は，以下の例文 2 のように引用を表わすために多く用いられる．例文 2 では，引用節の終端を越えて係る文節はないので，「本当に」が始端となる．

（例文 2）

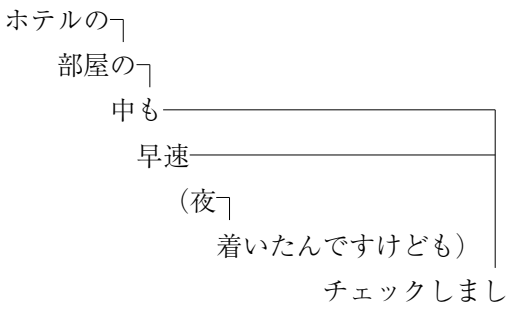


以降，引用節・トイウ節を合わせて引用節とする．

（挿入節）

挿入節は，発話の途中で話者の発話プランが変更されたとき，節の途中に別の節が注釈のような形で挿入されることにより発生するものである．書き言葉ではこのような表現はあまり用いられない．例文 3 では，() 内が挿入節に相当する．

（例文 3）



CSJ では，挿入節の終端の文節に「挿入節」というラベルが付与されている．挿入節の終端は基本的に強境界となっているが，挿入節を越えて前方から後方に係る係り受けが存在するため，文境界ではなく挿入節の終端と認定される．

2.2 節境界の曖昧さが係り受け解析に及ぼす影響

従来研究では，話し言葉において節境界の曖昧さが係り受け解析に及ぼす影響については，ほとんど考慮されていなかった．下岡ら (下岡, 内元, 河原, 井佐原 2005) は，話し言葉では文境界が曖昧であることが係り受け解析に与える影響が最も大きいことを指摘し，その影響を定量的に示した．彼らは，正しい文境界の情報を与えることにより，文境界を自動推定した場合に比べて約 3%高い係り受け解析精度が得られると報告している．また，文境界を推定する方法および文境界の自動推定結果を係り受け解析に利用する方法を提案し，その有効性も示した．しかし，その他の節境界については，係り受け解析に及ぼす影響は明らかではなかった．大野ら

(Ohno et al. 2006) は，文を節境界で分割して得られる節境界単位を基本として，節境界単位内の係り受けと節境界単位間の係り受けを別々に解析する方法を提案し，その有効性を示している．しかし，節境界単位は節とは異なるため，本来は節を超える係り受けを正しく推定することができない．例えば，2.1 の例文 1 や例文 3 では，節の始端は節境界ではないため，「昔から」と「思っていたところです」，「早速」と「チェックしました」は節境界単位をまたぐ係り受けとみなされ，正しく推定することができない．内元ら (Uchimoto et al. 2006) は文境界，言い直しの存在，挿入節・引用節などの境界の曖昧さ，係り先のない文節に着目し，正しい文境界の情報を与えた場合，さらに言い直し関係のうち係り元の文節を削除した場合，さらに挿入節・引用節の境界の情報を与えた場合，さらに係り先のない文節を削除した場合のそれぞれについて，係り受けモデルを学習しテストした場合に得られる係り受け解析精度を調べた．その結果，挿入節・引用節の境界の情報を与えた場合に約 2%高い精度が得られたと報告している．これは，話し言葉においては引用節や挿入節を含む文は節構造が複雑で，引用節あるいは挿入節の内部と外部とを結んでしまう係り受け解析誤りが多くなるためであると考えられる．逆に，引用節・挿入節の範囲を取得することができれば，係り受け解析精度の向上が期待できるが，そこまでは明らかにはされていない．そこで，本論文では，引用節・挿入節を自動認定する手法，および，その結果を利用して係り受け解析を行なう手法を提案し，引用節・挿入節を自動認定した結果を用いることで係り受け解析精度が有意に向上することを示す．手法については，3 章で詳しく述べる．

2.3　係り受け解析におけるその他の話し言葉特有の現象の扱い

その他の話し言葉特有の現象および本研究における係り受け解析時の扱いについては次の通りである．

(1) 文境界が明示されていない

話し言葉では文境界が明示されない．そのため，すべての文節に対して係り受けを特定しようとすると，文間関係も文節の関係として特定することになる．しかし，文間関係については人間の判断が揺れる場合が多い．また，自動要約のために文圧縮をしたり格関係を抽出する場合など，実際に必要となる係り受けの情報は文単位の係り受けであることが多い．そこで本研究では，文間関係は推定せず文境界を推定するにとどめ，係り受けは文内の文節間係り受けのみを対象として解析する．

(2) 係り先がない文節がある

話し言葉では，途中で発話のプランが変わったために係り先が消失したり，またフィラーや言いよどみなど，係り受け関係を特定しても用途がほとんど考えられず，係り受けを定義することに意味がない場合がある．このような場合，CSJ では係り受けが付与されていない．フィラーや言いよどみについては，浅原らの手法 (浅原 松本 2003) を用いるこ

とである程度特定できると考え，本研究ではすべて削除して扱う．ただし，どこにフィラーがあったかについての情報は残しておき，後の解析に利用する．本来これらの文節については，正しく「係り先なし」と推定するべきであるが，これについては今後の課題とする．それ以外の係り先を持たない文節については，以下に述べる条件に従って便宜的に係り先を設定する．

- 挿入節の終端の文節は，交差を発生させない範囲で文内のできるだけ後方に係るとする．2.1 節の例文 3 では，「着いたんですけども」の係り先は「チェックしました」とする．
- 引用節や挿入節の内部に絶対境界・強境界が含まれる場合，その内部境界の直前の文節の係り先は，後方に最初に現れる内部境界の直前または引用節・挿入節の終端の文節とする．以下の例文 4 にその例を示す．「：」は内部境界を表わす．「必要かな」「確保できないし」は係り先を持たないが，それぞれ「確保できないし」「作れるんじゃないかな」に係るとする．

  （例文 4）

  ｛やっぱり　ナイフは　必要かな：
  　ナイフが　ないと　何も　確保できないし：
  　　まずは　もしかしたら　何年間も
  　　掛けて　カヌーぐらい　作れんじゃないかな｝と　思いましてね

- 上記以外の係り先を持たない文節は，直後の文節に係るとする．

(3) 係り受け関係が交差する

一般に，日本語の書き言葉においては「係り受け関係は互いに交差しない」という非交差条件が成り立つと言われている．しかし，話し言葉ではこの非交差条件が成り立たないことも多い．例えば，以下の例文 5 では，「これが」が「正しいと」に係り，「私は」が「思う」に係るので係り受け関係が交差している．

（例文 5）

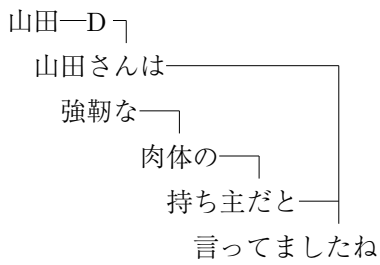
これが
　私は
　　正しいと
　　思う

しかし，今回用いた 188 講演において，係り受け関係が交差している箇所は 689 個とそれほど多くないため，本論文では，係り受けの非交差条件が成り立つと仮定して係り受け解析を行なう．したがって，評価の際，交差している係り受けのいずれかは解析誤りとなる．交差している場合への対処については今後の課題である．

(4) 言い直しが多い

話し言葉ではしばしば言い直しが生じる．CSJ では言い直し関係には係り受け関係と同様の関係が付与され，さらに D というラベルが付与されている．以下の例文 6 にその例を示す．

（例文 6）

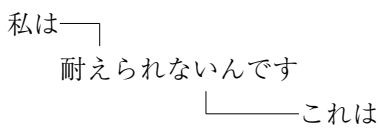


本来は，文節間の関係の推定のみではなくそれがどういった関係なのかまで推定すべきである．しかし，書き言葉を対象にした研究においても多くの場合は関係の有無の推定のみを対象としているため，本論文でも同様に，言い直し関係を係り受け関係として特定し，言い直し関係かどうかのラベルの推定までは行なわない．

(5) 倒置表現がある

話し言葉ではしばしば倒置表現が用いられる．CSJ では，倒置は左係りで表現されている．本論文では，関係を特定することが重要と考え，CSJ における倒置に対しては修正を行ない，便宜上すべて右係りとして扱った．例えば，以下の例文 3 では，「これは」が「耐えられないんです」に倒置で係っているが，「耐えられないんです」が「これは」に係るように修正した．

（例文 7）



なお，上記の対処法については (2) 以外は下岡らの手法 (下岡他 2005) に従っている．

3　係り受け解析と引用節・挿入節の自動認定のアプローチ

3.1　係り受け解析と境界推定の相互処理

図 1 に本手法で提案する処理の概要を示す．処理の流れは下記の通りである．入力は，形態素および文節の情報が付与されたテキストであり，CSJ を対象とする場合，一講演のテキストおよび形態素，文節の情報が入力となる．図 1 およびその説明において，文境界と引用節・挿入節の境界をまとめて境界と表現している．

- 境界推定（1 回目）

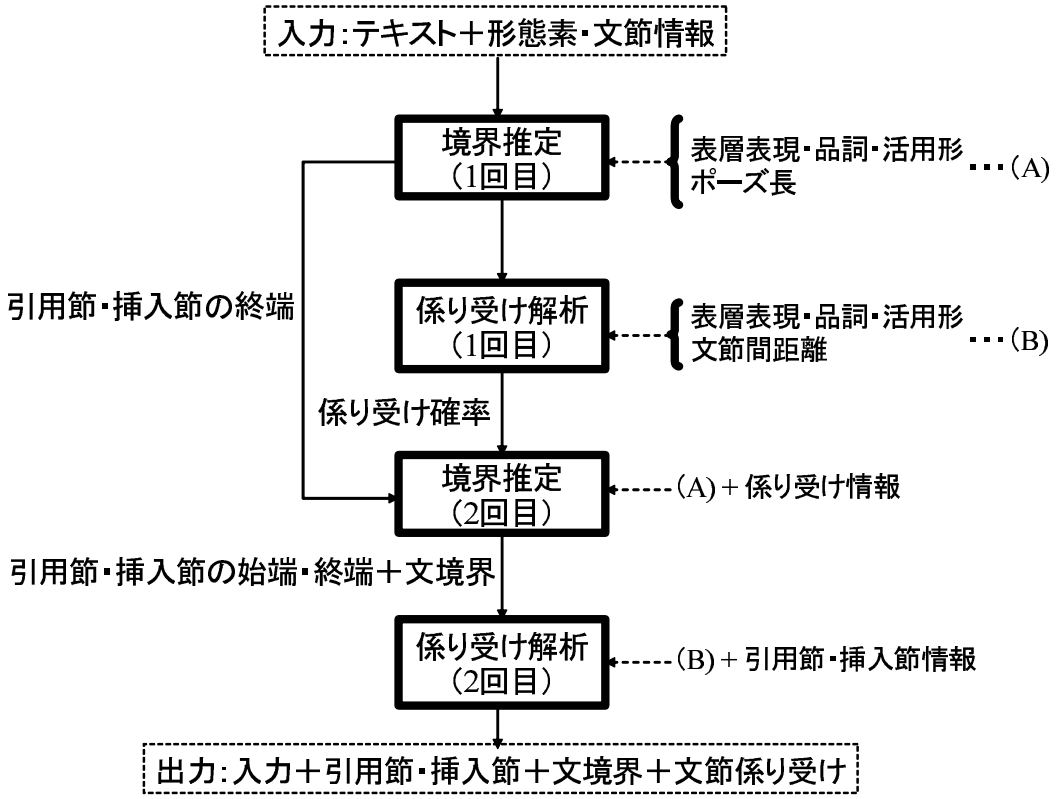


図 1　係り受け解析と境界推定の相互処理の概要

入力テキストに対し，まず，3.3 節で述べる手法により，表層表現・品詞・活用形，ポーズ長の情報などを素性として用いて，文境界，引用節，挿入節の境界を推定する．このとき，引用節・挿入節および文境界の 3 つの境界の推定は同時に行なう．

- 係り受け解析（1 回目）
  次に，境界推定（1 回目）で推定された文境界によりテキストを文に分割し，各文について，3.2 節で述べる手法により係り受け解析を行なう．このとき，素性としては表層表現・品詞・活用形，文節間距離などを用いる．境界推定（1 回目）で得られた情報のうち，引用節・挿入節の境界に関する情報はここでは用いない．
- 境界推定（2 回目）
  さらに，元の入力テキストに対し，文境界，引用節，挿入節の境界を再推定する．このとき，係り受け解析（1 回目）で得られた係り受けの確率の情報も素性として用いる．この素性は境界の情報により場合分けされており，その場合分けには，境界推定（1 回目）で得られた引用節・挿入節の境界のうち，終端の情報を用いる．
- 係り受け解析（2 回目）

最後に，境界推定（2 回目）で得られた文境界により元の入力テキストを文に分割し，各文について，3.2 節で述べる手法により係り受けの再解析を行なう．このとき，境界推定（2 回目）で得られた引用節・挿入節の境界の情報も素性として用いる．

以上の処理により，入力テキストに対し，文境界，引用節，挿入節の境界情報，および，各文内について文節係り受けの情報が得られる．

以下では，係り受け解析および引用節・挿入節の自動認定の手順についてそれぞれ説明する．

3.2 係り受け解析

本研究では，内元らの手法 (内元他 2000) に基づき，係り受け解析モデルを統計的に学習する．統計的係り受け解析では，文中の各文節がどの文節に係りやすいかを確率値で表わし，それらを要素とした係り受け行列を作成する．そして，一文全体が最適な係り受け関係になるように，それぞれの係り受けを決定する．ここで，2 つの文節間の関係を「間」「係る」「越える」の 3 カテゴリとして学習することにより，着目している 2 文節の間にある文節や，それらより後方にある文節との関係も考慮して確率値を計算できる．この係り受け解析モデルは最大エントロピー (ME) モデルとして実装され，素性には，単語の表層表現・品詞・活用形・文節間距離など，およびそれらの組合せが利用されている．

本研究ではさらに，着目している 2 文節の係り受けを仮定した場合に，その係り受けが引用節・挿入節の境界と交差するかどうかを素性に加える．より具体的には次の通りである．仮定した係り受けと引用節・挿入節の境界との関係は下記の 3 つの場合に分類できる．そして，引用節と挿入節のそれぞれについて，2 文節の関係が下記の分類のうちどれに属するかを素性値として与える．

- **仮定した係り受けと引用節・挿入節の境界とが交差する場合**
  交差が発生するのは，以下の 2 通りの場合である．このとき，2 文節が実際に係り受け関係を持つことはない．ただし，対象の 2 文節のうち係り文節が引用節あるいは挿入節の終端となっている場合はこの分類に含めない．
  - 2 文節の一方のみが引用節・挿入節の内部に含まれる
  - 2 文節の双方が異なる引用節・挿入節の内部に含まれる
- **仮定した係り受けと引用節・挿入節の境界とが交差しない場合**
  交差が発生しないのは，以下の 2 通りの場合である．
  - 2 文節がともに引用節・挿入節の内部に含まれない
  - 2 文節がともに同一の引用節・挿入節の内部に含まれる
- **2 文節のうち係り文節が引用節・挿入節の終端となっている場合**
  この場合には，2.3 節で述べたように，節の外部と内部との係り受けが例外的に結ばれるので，別の分類とする．

ただし，引用節・挿入節の境界が適切に推定されていることが望ましいため，この素性は図 1 の係り受け解析（2 回目）のみに用いる．この素性を用いることにより，2 文節間に仮定した係り受けと引用節・挿入節の境界とが交差する場合には，この 2 文節が実際に係り受け関係を持つ確率は低く推定される．

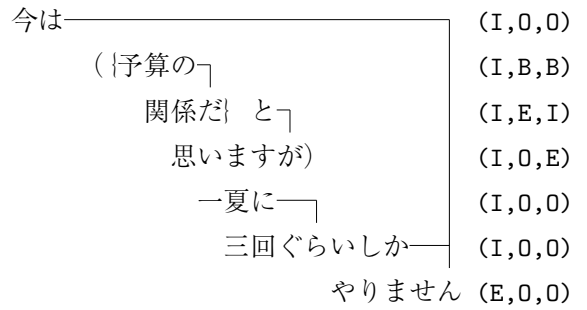
3.3　引用節・挿入節の自動認定

本研究では，下岡らが提案した機械学習による文境界推定法 (下岡他 2005) に基づき，引用節・挿入節の自動認定をテキストチャンキングの問題として扱う．これにより，引用節・挿入節の自動認定と文境界推定を同時に行なうことが可能となり，これらを別々に行なう場合に比べて，文境界推定の誤りに対しても頑健に動作することが期待できる．

テキストチャンカには，SVM (Support Vector Machines) に基づく YamCha(Kudo and Matsumoto 2001) を用いる．YamCha では，カーネル関数として多項式カーネルを用いることにより，複数の素性の組合せを考慮した学習が可能である．また，推定により得られた前後のチャンクラベルを動的素性として用いることができる．

本手法では，チャンクラベルは文節ごとに付与する．ラベルには，文境界に関するタグ（E: 文末，I: 文末以外）と，引用節および挿入節に関するタグ（表 1）の 3 つ組を用いる．以下の例文 8 にラベル付与の例を示す．ラベル内のタグは，順に（文境界に関するタグ，引用節に関するタグ，挿入節に関するタグ）を表わしている．例えば「予算の」に付与されているラベル (I, B, B) は，この文節が文末の文節ではなく，引用節・挿入節の始端となっていることを示す．3 つのタグは同時に推定されるため，このモデルでは文境界・引用節・挿入節の関係が考慮されている．例えば，引用節・挿入節の範囲が文境界を越えることはないので，(E, I, O) などというラベルが推定されることはない．

（例文 8）



YamCha の多項式カーネル次数は 3，解析方向は Right to Left とし，後方 3 文節の動的素性を利用する．SVM に与える素性としては，以下のものを用いる．

(1) 単語情報

単語情報として，表層表現・読み・品詞情報・活用の種類・活用形を用いる．引用節の

表 1　チャンキングに使用するタグの種類

タグ	タグの説明
B	引用節／挿入節の始端
E	引用節／挿入節の終端
I	引用節／挿入節の内部（始端，終端以外）
O	引用節／挿入節の外部
S	1 文節から成る引用節／挿入節

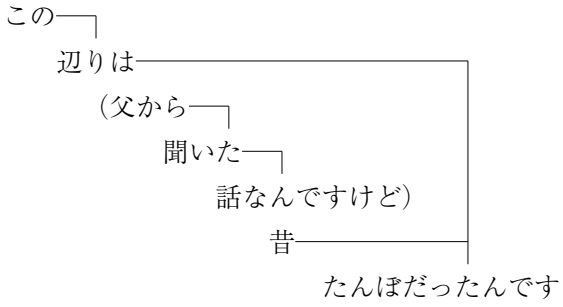
終端では「〜と思う」「〜って言う」などの表現が，挿入節の終端では「〜ですが」「〜けれども」などの表現が多用される．

(2) 文節の前後のポーズ長

引用節や挿入節の前後にはポーズが入りやすいと考えられる．そこで，文節の前後のポーズ長を素性として利用する．なおポーズ長としては，講演ごとに平均と分散で正規化した値を用いる．CSJ では，200 msec 以上のポーズで区切られた単位を転記単位として，書き起こしデータが作成されており，各転記単位には開始・終了時刻が付与されているため，これからポーズ長が計算できる．

引用節・挿入節の終端を推定する際には単語情報が大きな手がかりとなるが，以上の素性はすべて局所的な情報であり，これらだけから始端も同時に推定するのは困難である．例えば，以下の例文 9 では，「この辺りは父から聞いた話なんですけど」の部分だけを見た場合，「（他に自分が体験したことを話している途中で）この辺り（の話）は父から聞いた話なんですけど」という意味でも解釈できるため，「父から」が引用節の始端であるとは決定できない．この場合，「この辺りは父から聞いた話なんですけど」の全体が挿入節に含まれる可能性もある．

（例文 9）



このように，引用節・挿入節の始端を決定するためには，大域的な情報も必要となる．そこで，始端を決定する際には，自動推定した係り受けの情報をあわせて利用する．引用節・挿入節の終端が既に得られている場合，2.1 節および 3.2 節で述べたような引用節・挿入節と係り受け構造との関係により，始端より前の文節の係り受けには図 2 のような制約が成り立つ．本手

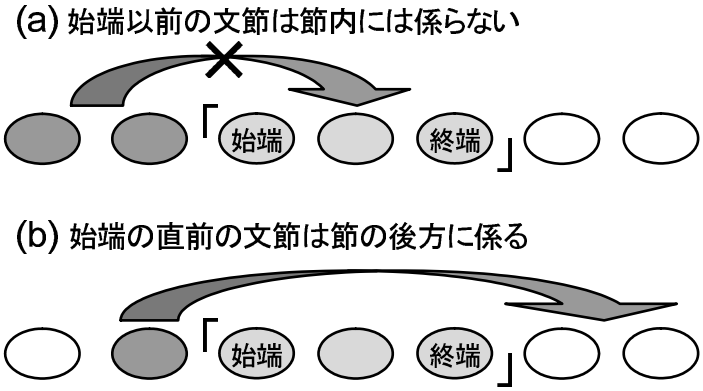


図 2 引用節・挿入節の始端以前の係り受けに関する制約

法ではこの制約を利用し，チャンキングを 2 回にわたって行なう．1 回目のチャンキング（図 1 の境界推定（1 回目））では，上述の素性のみを用いて文境界および引用節・挿入節を自動認定する．そして，ここで得られた文ごとに係り受け解析（図 1 の係り受け解析（1 回目））を行ない，1 回目のチャンキングで自動認定された引用節・挿入節の終端の情報をもとに，以下の係り受けの確率を素性に加えて，2 回目のチャンキング（図 1 の境界推定（2 回目））を行なう．学習データに対する係り受け確率は，学習データ内で 10-fold cross validation によって係り受け解析を行なうことで求める．

(a) 着目している文節より前方にある文節が，着目している文節と終端の間の文節に係る確率
(b) 着目している文節の直前の文節が，終端より後方の文節に係る確率

図 2 から，例えば，(a) の確率が小さく (b) の確率が大きければ，その文節は始端になりやすいと推測される．先の例文 9 では，「辺りは」「聞いた」「話なんですけど」は前方の文節が着目している文節に係るため，(a) の確率が大きくなる．また「父から」については，直前の文節「辺りは」が挿入節の終端「話なんですけど」より後方に係るため，(b) の確率が大きくなる．これより，「父から」が挿入節の始端であると推定できることが期待される．

4 評価実験

引用節・挿入節の自動認定および係り受け解析の評価実験を行なった．実験に用いたコーパスは CSJ のコア 188 講演（模擬講演 111 講演と学会講演 77 講演）の書き起こしである．この中には 6,148 個の引用節と 818 個の挿入節が含まれている．このうち 168 講演を学習データ，20 講演（模擬講演 11 講演と学会講演 9 講演）をテストデータとして用いた．

まず，下岡らの手法 (下岡他 2005) に従い，単語情報とポーズ長を用いて文境界を推定した後で，得られた文ごとに係り受け解析を行ない，ベースライン精度を求めた．文境界推定の F 値は 85.6 で，係り受け解析精度は，open テストで 77.7%, closed テストで 86.6% であった．closed テストでは，188 講演のすべてを学習に利用している．

4.1 引用節・挿入節の自動認定結果

3.3 節で述べた手法を用いて，引用節・挿入節の自動認定を行った．その結果を表 2 に示す．表 2 には以下の 5 種類の実験結果を示している．

- 係り受けを用いない場合（1 回目のチャンキング：図 1 の境界推定（1 回目））の認定精度
- open テストで得られた係り受けを用いた場合（2 回目のチャンキング：図 1 の境界推定（2 回目））の認定精度
- closed テストで得られた係り受けを用いた場合（2 回目のチャンキング：図 1 の境界推定（2 回目）の認定精度
- 正解の係り受けを用いた場合（2 回目のチャンキング：図 1 の境界推定（2 回目））の認定精度（係り受け確率はすべて 1.0 とする）
- 1 回目のチャンキング（図 1 の境界推定（1 回目））における終端のみについての認定精度

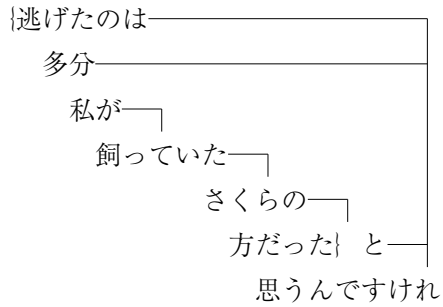
表 2 によると，引用節の終端のおよそ 9 割は正しく検出できている．検出できなかったものの中には「〜と」で終わる文末や，「〜っちゅう」「〜みたいな」など，使われる頻度が比較的少ない表層表現があった．始端とともに正解した精度は，open テストで自動推定された係り受けを利用することによって向上した．個々の文節における引用節のチャンクタグの推定結果についてマクネマー検定を行なったところ，$p < 0.01$ で有意な改善が得られていることが分かった．これは，本手法で素性として利用した係り受け情報が有効に作用したことを表わしている．

表 2　引用節・挿入節の認定精度（文境界が未知の場合）

	引用節			挿入節		
	再現率	適合率	F 値	再現率	適合率	F 値
係り受けを利用しない	40.6% (258/635)	44.6% (258/579)	42.5	2.6% (2/76)	25.0% (2/8)	4.8
係り受けを利用 (open)	42.2% (268/635)	46.0% (268/582)	44.0	2.6% (2/76)	14.3% (2/14)	4.4
係り受けを利用 (closed)	52.0% (330/635)	56.4% (330/585)	54.1	2.6% (2/76)	15.4% (2/13)	4.5
係り受けを利用（正解）	77.6% (493/635)	84.1% (493/586)	80.8	2.6% (2/76)	20.0% (2/10)	4.7
終端のみ一致	87.9% (558/635)	96.4% (558/579)	91.9	3.9% (3/76)	37.5% (3/8)	7.1

例えば，以下の例文 10 では，1 回目のチャンキングでは「多分私が飼っていたさくらの方だった」の部分が引用節だと誤って自動認定されたものの，2 回目のチャンキングで係り受けを利用することにより，「逃げたのは多分私が飼っていたさくらの方だった」の範囲が引用節であると正しく自動認定されるようになった．

（例文 10）



さらに，closed テストで得られた係り受けや正解の係り受けを用いた場合は，引用節の認定精度は大きく向上している．このことから，係り受け解析精度が改善されるのに伴って，引用節の認定精度も向上することが分かる．

一方，挿入節については，係り受けを利用してもほとんど検出できず，挿入節の終端の大半は文境界であると推定されていた．挿入節は，文末表現としてもよく用いられる「〜けれども」「〜ですが」の形で終わるものが多く，文境界との区別が難しいことが原因であると考えられる．これらの区別は，本手法で用いた素性だけでは困難である．そこで，4.5 節に述べるように，フィラーの有無や話速，韻律情報などを素性として用いてみたが，有意な精度向上は見られなかった．今後，より広範な素性を検討する必要があると考える．

4.2　節の自動認定結果を用いた係り受け解析結果

次に，自動認定された引用節・挿入節を用いて，3.2 節の手法で係り受け解析（図 1 の係り受け解析（2 回目））を行なったところ，表 3 に示す結果となった．ここで用いる引用節・挿入節の自動認定結果は，表 2 において open テストで得られた係り受けを利用したものである．学習データにおいても同様に，2-fold cross validation によって引用節・挿入節の自動認定を行なった．引用節・挿入節の自動認定結果を利用することで，open テストにおける係り受け解析精度

表 3　係り受け解析精度（文境界が未知の場合）

	open	closed
引用節・挿入節を利用しない	77.7%	86.6%
引用節・挿入節 (推定結果) を利用する	78.7%	86.9%
引用節・挿入節 (正解) を利用する	79.6%	87.7%

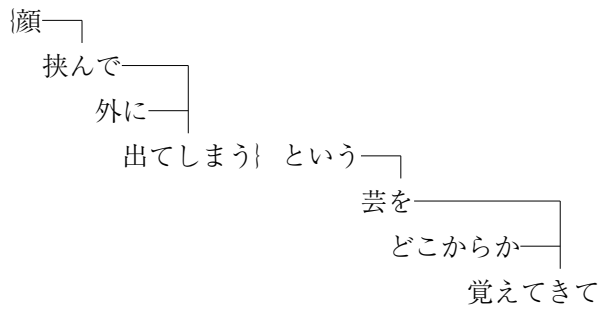
表 4　引用節・挿入節の境界と交差する係り受けの数（文境界が未知の場合）

		引用節		挿入節		誤り総数
		外→内	内→外	外→内	内→外	
引用節・挿入節を利用しない	open	302	217	117	3	639
	closed	226	69	114	2	411
引用節・挿入節（推定結果）を利用	open	322	128	121	1	572
	closed	261	50	112	1	424
引用節・挿入節（正解）を利用	open	136	22	78	1	237
	closed	84	4	78	0	166

は 1.0% 向上した．マクネマー検定を行なったところ，本手法を用いた係り受け解析精度はベースラインの精度より $p < 0.01$ で有意に上回っていることがわかった．この結果は，引用節・挿入節の推定に誤りがある場合でも，係り受け解析モデルが頑健に作用したことを示唆している．

そこで次に，引用節・挿入節を含む文の係り受け解析における解析誤りの数の変化について考察した．表 4 に，引用節・挿入節の内部と外部を結ぶ誤った係り受けが推定された数を示す．このような係り受け解析誤りの数は，引用節・挿入節の推定結果を利用することで，639 個から 572 個に削減された．特に，引用節の内部から外部へと係る解析誤りの数が，217 個から 128 個へと大きく削減された．その理由は次のように考えられる．一般に，引用節や挿入節がある場合は，その前方にある文節は引用節や挿入節を越えて遠くの文節に係ることが多い．その結果，従来の係り受けモデルでは，遠くに係る係り受けが誤って優先され，引用節・挿入節の内部から終端を越えて節の後方に係るような係り受け解析の誤りが多く発生していた．しかし，本手法によって，引用節については，認定精度の高かった終端の情報を活用することで，このような解析誤りを削減することができるようになったと考えられる．例えば，以下の例文 11 では，引用節・挿入節の情報を利用せずに係り受け解析を行なった場合には，「挟んで」（引用節内部）が「覚えてきて」（引用節外部）に係ると誤って推定されていたものの，「顔挟んで外に出てしまう」の部分を引用節として自動認定できたことにより，「挟んで」（引用節内部）が「出てしまうという」（引用節内部）に係るように修正された．

（例文 11）



また，表 3 には，引用節・挿入節の正解を与えた場合，すなわち認定精度が 100% だったと仮定した場合の係り受け解析の結果も示す．この場合，係り受け解析精度はさらに改善されており，引用節・挿入節の認定精度の向上に伴って係り受け解析精度も改善されることが分かる．

4.3 節の自動認定と係り受け解析の相互作用に関する考察

上述の実験結果から，引用節・挿入節の自動認定および係り受け解析の精度は，相互の情報を利用することにより高精度化されることが確認できた．単純には，同様のサイクルを繰り返すことにより，さらなる精度向上が期待される．そこで，引用節・挿入節の自動認定結果と係り受け解析の結果を再度相互に利用して，それぞれの精度がさらに改善されるかどうか調べた．しかしながら，引用節の認定精度および係り受け解析精度に有意な変化は見られなかった．これは，一度引用節・挿入節の情報を利用して推定した係り受けは，現在得られている節の認定範囲に対して最適な状態になっており，その結果を用いても始端の位置はほとんど修正できないためと考えられる．

逆に，再度相互に推定結果を利用することで，引用節の外部から内部へと係る解析誤りがわずかに増加する結果となった．これは，2 回目のチャンキングで引用節の始端を再推定する際に，誤った係り受けの情報が優先され，始端の位置が誤って文頭側にずれたことが原因と推測される．今後の課題として，特に引用節の始端付近について係り受けの傾向を詳細に分析し，より適切な係り受けの利用法を検討したい．

4.4 文境界が既知の場合の実験結果

次に，文境界推定の誤りの影響を調べるために，正解の文境界を与えて，引用節・挿入節の自動認定および係り受け解析を行なった．評価結果を表 5 ～ 表 7 に示す．

表 5 引用節・挿入節の認定精度（文境界が既知の場合）

	引用節			挿入節		
	再現率	適合率	F 値	再現率	適合率	F 値
係り受けを利用しない	50.6% (321/635)	53.7% (321/598)	52.1	18.4% (14/76)	25.9% (14/54)	21.5
係り受けを利用 (open)	51.5% (327/635)	54.9% (327/596)	53.1	23.7% (18/76)	30.0% (18/60)	26.5
係り受けを利用 (closed)	63.1% (401/635)	67.2% (401/597)	65.1	21.1% (16/76)	28.6% (16/56)	24.2
係り受けを利用（正解）	84.6% (537/635)	91.3% (537/588)	87.8	43.4% (33/76)	61.1% (33/54)	50.8
終端のみ一致	90.7% (576/635)	96.3% (576/598)	93.4	48.7% (37/76)	68.5% (37/54)	56.9

表 6　係り受け解析精度（文境界が既知の場合）

	open	closed
引用節・挿入節を利用しない	80.8%	90.3%
引用節・挿入節（推定結果）を利用する	81.4%	90.3%
引用節・挿入節（正解）を利用する	82.5%	91.3%

表 7　引用節・挿入節の境界と交差する係り受けの数（文境界が既知の場合）

		引用節		挿入節		誤り総数
		外→内	内→外	外→内	内→外	
引用節・挿入節を利用しない	open	250	201	86	6	543
	closed	179	58	82	3	322
引用節・挿入節（推定結果）を利用	open	289	113	82	7	491
	closed	218	44	66	3	331
引用節・挿入節（正解）を利用	open	112	24	46	1	183
	closed	55	2	32	0	89

結果として，文境界を与えることにより，引用節・挿入節の認定精度，係り受け解析精度ともに大きく上昇した．表 5 からは，引用節だけでなく挿入節についても係り受けを利用することで認定精度が向上すること，表 6 からは，引用節・挿入節の自動認定結果を用いることで open テストでの係り受け解析精度が 0.6% 向上することなどが分かる．また，引用節・挿入節の正解を与えた場合，係り受け解析精度はさらに改善されることも分かる．これらの結果は，文境界推定の誤りの影響がいかに大きいかを示している．

しかしながら，話し言葉において曖昧となる引用節・挿入節および文境界の情報をすべて与えても，書き言葉における係り受け解析精度と比べると依然として大きな差がみられる．話し言葉における係り受け解析精度をさらに向上させるためには，話し言葉特有の問題点について，さらに調査を行なう必要がある．これは今後の課題である．

4.5　その他の素性を追加した場合の実験結果

3.3 節で述べた素性 (1) と (2) に下記の素性を加え，それぞれの素性の組み合わせを用いて 4.1 節や 4.2 節と同様の実験を行なった．

文節の前後のフィラーの有無

引用節や挿入節の前後にはポーズだけでなくフィラーも入りやすいと考えられる．そこで，文節の前後のフィラーの有無も素性として利用する．

文節の話速

挿入節では，話者が早口になると考えられるため，各文節の話速をポーズ長と同様に正規化してから用いる．話速は，モーラあたりの平均発声時間によって定義する．すなわ

ち文節 b の話速 $rate(b)$ は，文節 b が転記単位 u に含まれるとき，次式で計算できる．

$$rate(b) = \frac{t_{end}(u) - t_{begin}(u)}{mora(u)}$$

ここで $t_{begin}(u)$, $t_{end}(u)$ は転記単位 u の開始・終了時刻を表わし，$mora(u)$ は転記単位 u に含まれるモーラ数である．

文節内の基本周波数の最大値

引用節・挿入節の境界の前後では，基本周波数 (F0) の上昇や下降が起こることが予想される．そこで，各文節における基本周波数の最大値を講演ごとに正規化したものを素性として用いる．CSJ では，F0 曲線の頂点や，曲線の変化率が大きく変わる点（屈曲点）に対して，自動抽出された F0 値が付与されており，素性としてはその値を用いる．

文節の先頭・末尾の韻律ラベル

CSJ では，韻律の変化に関するラベリングが行なわれている．ラベリング体系には，日本語の韻律ラベリング法として従来用いられてきた J_ToBI (Japanese Tones and Break Indices) (Venditti 1995) を自発音声に適用するための拡張が施された X-JToBI (eXtended J_ToBI) (前川 菊池 2001) が用いられている．これらのラベルは，音声の基本周波数のパターンや，音韻の時間長変化によるリズムを考慮して定義されたものである．

引用節・挿入節の始端や終端では，これらの韻律特徴に変化が起こることが考えられる．そこで，各文節の先頭および末尾に付与されている X-JToBI のトーン層ラベルを素性として用いる．X-JToBI で定義されているトーン層ラベルの例を表 8 に示す．

それぞれの素性の組み合わせに対し，個々のチャンクラベルの推定結果についてマクネマー検定を行なったところ，単語情報とポーズ長以外の素性，すなわち，フィラーの有無・話速・基本周波数・韻律ラベルを用いても，有意水準 $p = 0.01$ とした場合，有意な改善は得られなかった．これは，話速・基本周波数・韻律ラベルといった音響的特徴の現れ方が，引用節・挿入節において不安定であることや，上記の素性から得られる情報がすでに単語情報やポーズ長から得られていることなどが原因と考えられる．

表 8 X-JToBI トーン層ラベルの例

ラベル	ラベルの説明
%L	アクセント句頭境界
L%	下降調の句末境界
H%	単純な上昇調の句末境界
LH%	低ピッチ区間を含む上昇調の句末境界
HL%	上昇下降調の句末境界
HLH%	上昇下降上昇調の句末境界

5 おわりに

本論文では，CSJ を対象として，引用節・挿入節を自動認定し，その自動認定結果を係り受け解析に適用する手法について述べた．評価実験により，自動認定した引用節・挿入節の情報を係り受け解析に利用することで，係り受け解析精度が改善されることを示した．特に，引用節の終端は高い精度で推定することができたため，その情報を利用することで，引用節の内部から終端を越えて外部に係る解析誤りを削減することができた．今後の課題としては，実験の考察を踏まえ，より広範な素性の考慮，より適切な係り受けの利用法の検討などにより，さらなる精度の改善を図ることや，音声認識結果に誤りがある場合の頑健性について検討することなどが挙げられる．また，係り受け解析における話し言葉特有の問題点についてもさらなる調査を行ないたい．

参考文献

浅原正幸, 松本裕治 (2003). “形態素解析とチャンキングの組み合わせによるフィラー／言い直し検出.” 言語処理学会 第 9 回年次大会 発表論文集, pp. 651–654.

藤尾正和, 松本裕治 (1997). “統計的手法を用いた係り受け解析.” 情報処理学会 自然言語処理研究会 NL117-12, pp. 83–90.

古井貞熙, 前川喜久雄, 井佐原均 (2000). “科学技術振興調整費開放的融合研究推進制度―大規模コーパスに基づく『話し言葉工学』の構築―.” 日本音響学会誌, **56** (11), 752–755.

春野雅彦, 白井諭, 大山芳史 (1998). “決定木を用いた日本語係受け解析.” 情報処理学会論文誌, **39** (12), 3177–3186.

Kawahara, D. and Kurohashi, S. (2006). “A Fully-Lexicalized Probabilistic Model for Japanese Syntact ic and Case Structure Analysis.” In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2006)*, pp. 176–183.

工藤拓, 松本裕治 (2002). “チャンキングの段階適用による係り受け解析.” 情報処理学会論文誌, **43** (6), 1834–1842.

工藤拓, 松本裕治 (2004). “相対的な係りやすさを考慮した日本語係り受け解析モデル.” 情報処理学会論文誌, **46** (4), 1082–1092.

Kudo, T. and Matsumoto, Y. (2000). “Japanese Dependency Structure Analysis Based on Support Vector Machines.” In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 18–25.

Kudo, T. and Matsumoto, Y. (2001). “Chunking with Support Vector Machines.” In *Proceedings*

of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2001), pp. 192–199.

黒橋禎夫, 長尾眞 (1994). “並列構造の検出に基づく長い日本語文の構文解析.” 自然言語処理, **1** (1), 35–57.

黒橋禎夫, 長尾眞 (1997). “京都大学テキストコーパス・プロジェクト.” 言語処理学会 第 3 回年次大会 発表論文集, pp. 115–118.

前川喜久雄, 菊池英明 (2001). “X-JToBI：自発音声の韻律ラベリングスキーム.” 電子情報通信学会技術研究報告, SP2001-106, pp. 25–30.

丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝 (2003). “節境界自動検出ルールの作成と評価.” 言語処理学会 第 9 回年次大会 発表論文集, pp. 517–520.

Matsubara, S., Murase, T., Kawaguchi, N. and Inagaki, Y. (2002). “Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language.” In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp. 640–645.

Ohno, T., Matsubara, S., Kashioka, H., Maruyama, T. and Inagaki, Y. (2006). “Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries.” In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 169–176.

白井諭, 池原悟, 横尾昭男, 木村淳子 (1995). “階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度.” 情報処理学会論文誌, **36** (10), 2353–2361.

下岡和也, 内元清貴, 河原達也, 井佐原均 (2005). “日本語話し言葉の係り受け解析と文境界推定の相互作用による高精度化.” 自然言語処理, **12** (3), 3–18.

高梨克也, 丸山岳彦, 内元清貴, 井佐原均 (2003). “話し言葉の文境界―CSJ コーパスにおける文境界の定義と半自動認定―.” 言語処理学会 第 9 回年次大会 発表論文集, pp. 521–524.

内元清貴, 関根聡, 井佐原均 (1999). “最大エントロピー法に基づくモデルを用いた日本語係り受け解析.” 情報処理学会論文誌, **40** (9), 3397–3407.

内元清貴, 村田真樹, 関根聡, 井佐原均 (2000). “後方文脈を考慮した係り受けモデル.” 自然言語処理, **7** (5), 3–17.

内元清貴, 丸山岳彦, 高梨克也, 井佐原均 (2003). “『日本語話し言葉コーパス』における係り受け構造付与.” 平成 15 年度国立国語研究所公開研究発表会予稿集.

Uchimoto, K., Hamabe, R., Maruyama, T., Takanashi, K., Kawahara, T. and Isahara, H. (2006). “Dependency-structure Annotation to Corpus of Spontaneous Japanese.” In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 635–638.

Venditti, J. (1995). “Japanese ToBI Labelling Guidelines.” *Papers from the Linguistics Labora-*

tory. Ohio State University Working Papers in Linguistics, **50**, 127–162.

略歴

浜辺　良二：2005 年京都大学工学部情報学科卒業．2007 年同大学院情報学研究科修士課程修了．現在，パナソニックコミュニケーションズ株式会社に勤務．在学中，話し言葉処理の研究に従事．

内元　清貴：1994 年京都大学工学部電気工学第二学科卒業．1996 年同大学院修士課程修了．博士（情報学）．同年郵政省通信総合研究所入所．現在，独立行政法人情報通信研究機構主任研究員．自然言語処理の研究に従事．言語処理学会，情報処理学会，ACL，各会員．

河原　達也：1987 年京都大学工学部情報工学科卒業．1989 年同大学院修士課程修了．1990 年同博士後期課程退学．同年京都大学工学部助手．1995 年同助教授．1998 年同大学情報学研究科助教授．2003 年同大学学術情報メディアセンター教授．現在に至る．この間，1995 年から 1996 年まで米国ベル研究所客員研究員．1998 年から ATR 客員研究員．1999 年から 2004 年まで国立国語研究所非常勤研究員．2001 年から 2005 年まで科学技術振興事業団さきがけ研究 21 研究者．音声言語処理，特に音声認識・理解に関する研究に従事．京大博士（工学）．1997 年度日本音響学会粟屋潔学術奨励賞受賞．2000 年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表，IEEE SPS Speech TC 委員，IEEE ASRU 2007 General Chair，言語処理学会理事，を歴任．情報処理学会音声言語情報処理研究会主査．日本音響学会，人工知能学会 各評議員．情報処理学会，電子情報通信学会，言語処理学会，IEEE 各会員．

井佐原　均：1978 年京都大学工学部電気工学第二学科卒業．1980 年同大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所．現在，独立行政法人情報通信研究機構上席研究員およびタイ自然言語ラボラトリー長．自然言語処理，語彙意味論の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．

（2007 年 3 月 22 日　受付）
（2007 年 7 月 4 日　再受付）
（2008 年 8 月 10 日　採録）